# Chapter 1

# Introduction

*AI has by now succeeded in doing essentially everything that requires 'thinking' but has failed to do most of what people and animals do 'without thinking' - that, somehow, is much harder!*

- Donald Knuth

The human race has long been fascinated with machines that have human-like intelligence. Intelligence, at the very least, is what allows us to perceive the environment, make sense of it, and act accordingly. Artificial Intelligence is the effort towards making machines exhibit such intelligence. The first steps towards artificial intelligence involved identification of tasks requiring intelligence and developing algorithms to solve them [1]. Solving puzzles, playing games, proving theorems, are a few examples of tasks that require some amount of intelligence. It turned out that these problems actually have algorithmic solutions that can be easily implemented in a machine. However, there are other kinds of tasks that humans perform everyday that do not have straightforward algorithmic solutions. Recognizing an object, understanding what a scene depicts, understanding a narrative from a video, are some such examples. Humans learn how to perform these tasks from experience. We are able to recognize an activity as *fighting* because we have seen *fights* in the past and have formed an idea of what happens during a *fight*. Machines can be made to exhibit such intelligence by designing algorithms that could learn patterns from examples. Machine learning is the field that studies such algorithms.

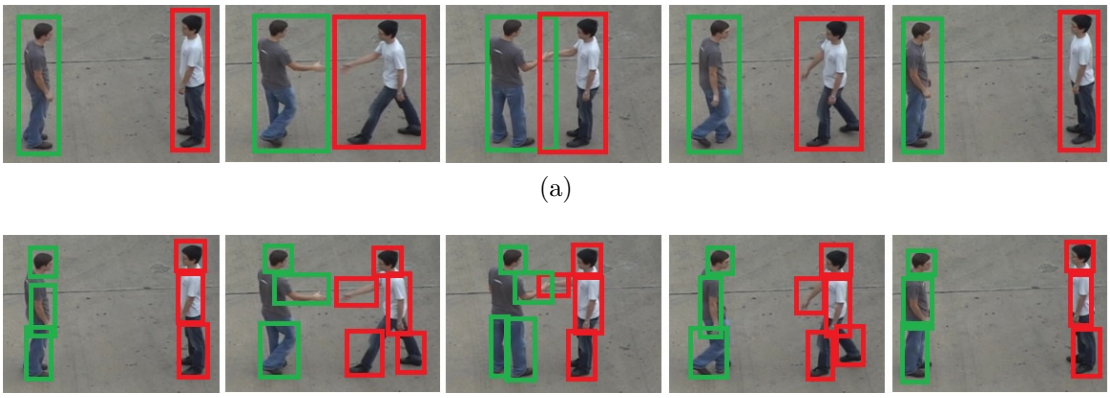
Learning algorithms for the kind of tasks discussed above come with a wide range of issues. Representation is one of the most basic issues; deciding how to best represent a picture or an activity sequence is fundamental to the problem of learning. For machines, a picture is stored as a matrix of quantitative pixel

values; a video is stored as a sequence of such pictures. However, a person understands a scene based on a *qualitative* rather than a quantitative description. A qualitative description of a scene may include identifying the spatial relations between the objects. This combined with an innate ability to find patterns has given us an intellectual advantage. However, extracting a qualitative description of any scene is far from trivial for a machine. As such, it is necessary to define procedures based on *geometric* computations to obtain a human-like qualitative description of the changing sequence of scenes during an activity. In this thesis, we explore *representations* and *reasoning procedures* that best capture a human-like abstraction and modelling for the problem of *human activity recognition* within a video.

## 1.1 Context

A video captures real world events occurring in space over time using a sequence of snapshots. Recognition of human activities within a video is a subject of great attention because of its applications in various interesting problems such as content-based retrieval in a video database, automated surveillance, automated patient monitoring systems and ambient assisted living. Human activity recognition (HAR) is concerned with correctly classifying input data into its underlying activity category [2]. The underlying activity category may be a verb, such as *handshaking* [3] or a complex activity, such as *making cereal* [4]. In this thesis, the input data is assumed to be in the form of a video and the activity category is a verb. HAR involves generation of activity models which are generalized descriptions of what constitutes an activity. These models are then used for detection and recognition of activities within the video [5, 6]. Fig. 1-1 shows an example of a *handshaking* activity from the UT Interaction dataset [3]. A generic model of the activity should describe how the two persons come close, touch one another, and move apart. An activity model with finer details may give the description as - two persons come close, *hold each others right hand*, and then move apart. Learning such an interaction model for the activity would allow the system to recognize the *handshaking* activity in any video thereafter. It stands to reason that an interaction model with finer details would be better at recognizing the activity.

Traditionally, activity analysis in video has been tackled using image processing and computer vision approaches [7]. Though considerable progress has been made using vision based approaches, it is seen that such techniques are susceptible to



(a)

(b)

Figure 1-1: *Handshaking* from the UT-Interaction Dataset

occlusions, noise, and are video dependent. With recent advances in technology, efforts have also been made towards extending such approaches to handle 3D data (particularly depth information) obtained through Kinect sensors [8, 9].

Most of the data-driven vision-based techniques while focusing on the pixel-information fail to take into account high-level contextual information contained in the video, such as the topological changes between different entities as an effect of the activity [10]. Qualitative Spatial Representation (QSR) is a computational apparatus that is inspired by the way humans describe and reason about space. Of late, QSR inspired approaches of HAR are gaining popularity [5, 11]. One of the primary reasons why QSR is used for the description of interactions is because visual information intuitively retained by a human being is qualitative [12]. Considering the innate, and compared to machines, far superior ability of human beings to reason about space and time, it makes sense to use a similar mechanism for description of space-time activities. Further, noise and video specific details of activities are easily abstracted away through a qualitative abstraction [13].

In most QSR approaches for HAR, humans are abstracted using single bounding boxes [11, 14]. However, abstracting the whole body using a single bounding box abstracts away a lot of important interaction details. Human interactions are better described when the human body is viewed as a collection of parts [15]; each part is abstracted through a bounding box. In Fig. 1-1a, the single bounding box for abstraction of the human bodies does not describe the *handshaking* activity in the required level of granularity. In Fig. 1-1b, when separate bounding boxes are used to abstract different parts of the human body, the qualitative description would include important interaction details involving the hands of the human bodies. However, a part-based model of the human body mostly views body parts as independent entities. This is counter-intuitive to the notion of body-parts being

*part* of a *whole* body. Herein lies the motivation for an efficient representation of human interaction that abstracts human bodies as collection of parts, while retaining the part-whole relation. In this thesis, such an abstraction is termed as *extended object*, which is similar to the notion of *composite regions* [16] or *disjoint regions* [17].

CORE9 is a comprehensive representation for encoding and extraction of several interesting spatial information, such as topology, direction, size, distance, and motion, between two axis-aligned rectangles [13]. To the best of our knowledge, a similar representation or reasoning framework for extended objects is not discussed in the literature. Therefore, in this thesis, a framework for extracting various qualitative spatial relations, viz. topology, direction, and distance, between extended objects is presented. This is the *Extended CORE9* framework.

Activities can be seen as temporal evolution of spatial relations between the interacting humans or between human and objects. This calls for an appropriate representation of activities that keeps track of the evolution of spatial relations between extended objects over time. To this end, a graph-based representation of activities, termed *Temporal Activity Graph* (TAG) is discussed. A generic solution to classification of graph representations involves computing similarities (or dissimilarities) between activities within existing analogy-based methods of classification, such as a Support Vector Machine (SVM). The thesis presents such an approach. A *Temporal Activity Graph Kernel* (TAG kernel) is defined to compute similarity between a pair of TAGs which is then used in an SVM for classification of activities.

Learning using graphs is well researched, but the TAG representation also contains temporal information. This necessitates investigations for appropriate learning mechanisms. In order to learn generic activity models that are capable of retaining high-level relational description of activities we present a grammar-based approach. A stochastic context free graph grammar, termed *Temporal Activity Graph Grammar* (TAG grammar), is presented. Further, a TAG grammar induction algorithm is proposed to learn the rules of the grammar. The TAG grammars model the activity classes, and can be used to recognize unseen activities using appropriate parsers. The thesis also discusses a parsing mechanism for recognizing activities represented as TAGs.

## 1.2 Objective

Early in the research, it was discovered that a proper abstraction of human bodies within a HAR system affects a large number of factors, including efficiency and effectiveness. This is achieved in this thesis by abstracting human bodies as *extended objects*. In the context of HAR in video, we define *extended objects* as a set of components, such that each component is approximated by an *axis-aligned minimum bounding rectangle*.

For a qualitative description of the human activities in a video, it is important to have an efficient mechanism for obtaining the spatial relations between the interacting entities. To this extent, we use geometric reasoning to reduce the amount of computation, thereby enhancing efficiency.

Further, an activity is spatio-temporal in nature. The spatial relations between the interacting entities evolve over time. A graph-based representation is capable of adequately encoding temporal information. However, the extended object based abstraction of entities necessitates investigations towards an appropriate graph representation. For the specialized graph-representation there is a need for suitable classification or learning mechanisms.

The objective of this thesis is to explore *qualitative and geometric reasoning for an extended object abstraction within graph representation of spatio-temporal activities to recognize human activities from video.*

## 1.3 Contributions

The first component of our contribution in this thesis is the **Extended CORE9** framework for efficiently computing qualitative relation between extended objects. We enumerate the ways in which CORE9 can be applied for an extended object based abstraction and the limitations of these approaches. Extended CORE9 includes an algorithm that uses geometric reasoning to efficiently compute qualitative relations between a pair of extended objects, pertaining to topology, direction, and distance. It is theoretically proved that for extended objects with at most $n$ components, the algorithm computes all concerned relations in $O(n\ log\ n)$ time on average. Further, it is shown via experiments that the set of qualitative relations obtained using Extended CORE9 provide a better classification of human activities. The experiments are conducted such that the qualitative relations thus obtained provide a bag-of-words description of the activities. The improvement in results using Extended CORE9 over CORE9 for HAR, even with a primitive bag-

of-words description, suggest that the extended object abstraction is promising.

In the bag-of-words description of activities, temporal information is not incorporated. Our next step is to present a representation of activities that incorporate spatial as well as temporal information. **Temporal Activity Graph** is presented for representation of activities, wherein interacting entities are abstracted as extended objects. For classification of activities represented as TAGs, a kernel based SVM classification is discussed. A **Temporal Activity Graph Kernel** is presented for computing similarity between a pair of TAGs. The TAG kernel computes similarity between TAGs by comparing similarity between the set of *label sequences* of the TAGs. A label sequence in a TAG is a sequence of qualitative spatial relations between the components of the extended objects. An *interestingness factor* is presented to reduce the number of comparisons necessary for comparing two sets. Similarity between label sequences are in turn computed based on the similarity between the qualitative relations. A novel *neighbourhood-based similarity* is presented in this thesis for computing similarity between a pair of qualitative relations. Experiments are conducted to show the effectiveness of the proposed kernel.

The TAG kernel based classification does not provide a model of the underlying structures of the activity classes. Therefore, a generative learning technique that models activities represented as TAGs using a probabilistic context free graph grammar is presented. Grammars are often used for encoding the recursive and hierarchical nature of human activities. To encode the structure of activities represented as TAGs using a grammar, a **Temporal Activity Graph Grammar** is proposed. Although generic graph grammar induction algorithms have been proposed in literature, they usually involve expensive computations that involve finding isomorphic subgraphs. A grammar induction algorithm is designed that incorporates the simplicity of string grammar induction algorithms by taking advantage of the uniform structure of TAGs. A modified parsing algorithm based on LR(0) string parser is then discussed for parsing activities presented as TAGs and modelled using TAG grammar. Experiments are conducted and the classification results using TAG grammar are compared to results reported in literature.

## 1.4 Thesis Outline

In Chapter 2, existing work related to this thesis are discussed along with the background knowledge necessary for a clearer understanding of the thesis. The chapter begins with an introduction to the field of HAR and its various applica-

tions. We attempt to give a rough idea of the breadth of work done with respect to the various aspects of the problem. The rising popularity of Knowledge Representation and Reasoning techniques for HAR in video is discussed followed by a gentle introduction to the field of Qualitative Spatial Representation and Reasoning. A discussion of the CORE9 framework and its variants along with the importance of extended object based abstraction with a qualitative representation is presented. Qualitative and geometric reasoning techniques are discussed thereafter. The is followed by a discussion on graph representation and the utility of a grammar based modelling.

In Chapter 3, Extended CORE9 is presented as a geometric reasoning framework for opportunistically computing qualitative relations for extended objects. An algorithm that computes qualitative spatial relations pertaining to various aspects of space for a pair of extended objects is presented. Results of experiments performed are reported that show how relations between humans abstracted as extended objects, obtained using Extended CORE9, give a better description of the activities.

In Chapter 4, representation of activities as TAGs is discussed. A TAG kernel is then presented to compute similarity between two activities represented as TAGs so that similarity based classification techniques can be used. Various measures to compute *neighbourhood-based similarity* of two qualitative relations, similarity between two edge labels, similarity between two label sequences are put together to compute the similarity between two TAGs. Further, two intrinsic orders on the label sequences - *skeletal information* and *interestingness* - are presented. The algorithm is theoretically analyzed to show that it can be adequately used as a kernel. Experiments are performed using the kernel function in a SVM classifier and results are discussed.

In Chapter 5, a TAG grammar for representing general descriptions of activity classes is presented. The TAG grammar is a probabilistic context free graph grammar. An algorithm for inducing rules of the TAG grammar is presented. A parsing mechanism is discussed for the TAG grammar that can be used for recognizing an activity represented as a TAG. Experiments are then performed to evaluate the effectiveness of such a grammar in recognition of human activities.

In Chapter 6, the conclusion of the thesis is drawn and possible future directions of work are presented. Possible extensions to all the key components presented in the thesis are discussed that include applications in single-person action recognition and firewall anomaly detection. The drawbacks of the work presented in this thesis and the possible directions of future work are pointed out.